On the Quality of Privacy Policy Documents of Virtual Personal Assistant Applications

Chuan Yan
University of Queensland
Australia

Fuman Xie
University of Queensland
Australia

Mark Huasong Meng
Institute for Infocomm Research
Singapore

Yanjun Zhang
University of Technology Sydney
Australia

Guangdong Bai*
University of Queensland
Australia

ABSTRACT

The app ecosystem built around virtual personal assistant (VPA) services has become flourishing in recent years. In response to the increasingly stringent data protection regulations, VPA service providers require application developers to include a privacy policy that declares their data handling practices. These privacy policies serve as the *de facto* agreement between developers and users, and may be taken as the basis for resolving conflicts in the event of a data breach. Therefore, it is essential that privacy policy documents are crafted in a clear, easy-to-understand, and unambiguous way.

In this work, we conduct the first systematic study on the quality of privacy policies in the VPA app domain. We identify four metrics that enable the quality of the privacy policy to become measurable, including *timeliness*, *availability*, *completeness*, and *readability*. We then develop QuPer, which extracts the meta features (e.g., update history) and linguistic features (e.g., sentence semantics) from privacy policies, and assesses their quality. Our analysis reveals that the status of the quality of privacy policies in the VPA app domain is concerning. For instance, only 1.17% of privacy policies completely cover all contents that are regarded as privacy concerns by legislation (e.g., GDPR Article 13) and relevant literature. Our findings are expected to raise an alert among the VPA app developers and provide them with guidelines for creating high-quality privacy policy documents.

KEYWORDS

privacy compliance, privacy policy, Alexa skills

1 INTRODUCTION

AI (artificial intelligence)-backed virtual personal assistant (VPA) services, such as Amazon Alexa [8] and Google Assistant [32], have gained tremendous popularity in recent years. Centered around them, an ecosystem similar to the one among mobile applications, which has proven a big success in the last decade, is growing rapidly. The VPA services enable third-party developers to create VPA applications (or *app*s for short), e.g., *skills* in Amazon Alexa and *actions* in Google Assistant, and release them through app stores. This allows users to easily enable and use the apps through their smart speakers. According to a recent report from Statista [77], VPA services have reached a global user base of billions.

The openness of this ecosystem raises a great privacy concern though. Dishonest VPA apps can appear in the app store, and once installed, can collect sensitive user information, such as location, name, age, and gender. Although current VPA services require the developers to declare the permissions [9] whenever their skills request to access personal data, some malicious apps can still bypass this and gather information at runtime, as revealed by recent studies [35, 81]. The privacy threats not only put users at risk, but also largely pose challenges for app developers and VPA service providers, particularly in the global context that many countries and regions have enacted stringent data protection legislation, e.g., the well-known European Union (EU) General Data Protection Regulation (GDPR) [40]. Any privacy breach can result in significant penalties to the data controllers and processors. For example, on July 16, 2021, Amazon was fined 746 million euros by Luxembourg's National Data Protection Commission for failing to comply with GDPR in protecting its users' data [10].

VPA services have indeed taken steps to mitigate privacy concerns. Taking Amazon Alexa as an example, skill developers are required to release a privacy policy document that discloses how their skills handle user data, including access, collection, use, and sharing of user data [10]. Nonetheless, the enforcement of this requirement remains problematic. A recent study [80] shows that out of 65,195 Alexa skills collected, only 21,063 skills provide privacy policies. Even though some skills do provide one, its quality is worrisome. Additional studies [16, 81, 82] find that many of the available privacy policies are incomplete, contain inaccessible resources or technical terms, or use a language that mismatches the skill's support languages. These issues greatly cause difficulties for users in understanding the privacy policies of the skills they use.

The privacy concerns of the VPA app ecosystem have also raised high attention from the research community. A line of research has been dedicated to detecting the runtime information gathering behaviors of VPA apps [35, 47, 79]. Some recent studies [81, 82] propose to examine the compliance between VPA apps' data handling behaviors and the statements in their privacy policies. However, the complementary problem of *what are the common problems of VPA apps' privacy policy documents and how to guide the VPA app developers to develop a high-quality privacy policy* remains open.

*Corresponding author.

Proceedings on Privacy Enhancing Technologies 2024(1), 478–493

https://doi.org/10.56553/popets-2024-0028

Our Work. In this work, we conduct the first systematic study to assess the quality of VPA apps' privacy policies. We first formulate a taxonomy to break down the quality of privacy policy documents into *measurable* metrics. To this end, we resort to two major data regulations, i.e., EU GDPR [40] and California Consumer Privacy Act (CCPA) [59], the standards working groups of European Data Protection Board (EDPB) that form privacy policy guidelines [27, 28], and the literature on privacy policies in VPA and other domains like mobile apps and websites [45, 50, 51, 57, 71, 74]. We summarize their quality concerns and propose four quality metrics, including *timeliness*, *availability*, *compliance of disclosure*, and *readability*.

Based on our taxonomy, we develop a framework named QuPer to assess the quality of the privacy policy, using machine learning and natural language process (NLP) techniques. It automatically collects and synthesizes *meta features* (e.g., update history) and *lingual features* (e.g., the semantics of sentences) of the given privacy policy. By analyzing these features, it can align the updates to the release and revision of an app (i.e., the timeliness), track the availability of the linked resources and required multilingual versions (i.e., the availability), assess the coverage of the contents required by legislation like GDPR Article 13 (i.e., the completeness), and evaluate its writing style (i.e., the readability).

We conduct a large-scale study with QuPer on all 65,195 available Amazon Alexa skills, the apps of the most popular VPA service, to understand the landscape of the quality of privacy policies among modern VPA apps. QuPer reveals that the current status of the privacy policy quality is concerning. Only 5,473 privacy policies are well formatted, and only 1.17% (64/5,473) have complete contents. The privacy policy documents refer to 200,183 external links, but 28,745 (14.3%) have become invalid. More than half of the analyzed privacy policies have issues for users to read and understand.

Contributions. The main contributions of this work are as follows.

- **Understanding the quality of the VPA app privacy policy**. We conduct the first comprehensive study on the quality of the VPA privacy policy. Our work proposes four quality metrics to measure different aspects from writing styles and temporal features, to contents and semantics.
- **A systematic approach of privacy policy quality measurement**. We develop QuPer that can automatically extract meta features and textual features of the privacy policy. It features a two-step document processing method that derives context-sensitive semantics from the sentences. This endows QuPer with the capability of inferring fine-grained information to determine content coverage and readability.
- **Revealing the *status quo* of privacy policy quality in existing VPA apps**. We study the landscape of privacy policy quality among Alexa skills. Our findings reveal that the current status of privacy policy quality remains concerning. Our work should raise an alert to VPA app developers, and encourage store operators to take actions for quality assurance. It can be extended for policy quality auditing in other domains.

2 A TAXONOMY OF QUALITY METRICS

To effectively assess the quality of privacy policies, it is necessary to establish a set of measurable metrics. In this section, we take the first step to construct a taxonomy of quality metrics. Since there is not a comprehensive list for both VPA and other domains, we resort to three sources to summarize the concerns on the privacy policy quality (Section 2.1), and based on them, we formulate the taxonomy for our assessment (Section 2.2).

2.1 Identifying Public Concerns on Privacy Policy Quality

2.1.1 Sources for identification. As user privacy protection becomes increasingly important, legislators, standard working groups, and the research community have made efforts to establish privacy policy guidelines or examine privacy-related documents. We thus turn to these sources to summarize their concerns regarding the quality of privacy policies, in order to create our taxonomy.

Data Regulations. We first turn to two major data regulations, i.e., the EU GDPR [40] and the California Consumer Privacy Act (CCPA) [59]. The GDPR is one of the pioneering comprehensive data protection laws that came into effect in 2018. It is designed to protect the personal data and privacy of individuals within the EU by imposing strict regulations on how organizations collect, store, process, and transfer their data. The CCPA is a state-level privacy law in California, United States that went into effect in 2020. It applies to businesses that collect, use, or disclose the personal information of California residents. While it has most principles in common with GDPR, it specifically grants California consumers the right to know what personal information businesses are collecting about them and to request that such information be deleted.

Guidelines. Besides data regulations that mainly provide high-level principles, we also consider guidelines from relevant working groups. In particular, we look into those provided by Article 29 Working Party [28] which is a team made up of 28 national regulators from multiple European Union countries. This team is dedicated to the protection of individuals with regard to the processing of their personal data, and aims to promote data protection by enforcing GDPR. In 2018, it was promoted into the EDPB (*European Data Protection Board*), gaining more legal weight to push through decisions.

We review the *Guidelines on transparency under Regulation 2016/679* document [27] from this working party, given that transparency is relevant to the GDPR principle of disclosing the data handling practices to the data owner. We mainly focus on the Articles 12 and 13 of this document, which are under "*Clear and plain language*" (pp.8 in [27]), as they present the requirements for the writing of privacy policies. More specifically, Article 12 requires that the privacy policy should be in clear and plain language. The information provided in the privacy policy should be simple and easy to understand, and the use of complex sentences and grammar should be avoided. Article 13 proposes that language qualifiers such as "may", "might", "some", "often", and "possible" should be avoided in the privacy policy. It also requires that when the data controller is targeting data subjects in one or more languages, a set of privacy policies in these respective languages should be provided.

Literature. We refer to the literature to recognize quality metrics from the concerns of the research community. We start with Micheti et al. [57] since this is an early study on the recommendations for drafting privacy policies (with a specific target of young people). It proposes the guidelines of the privacy policy writing based on

user studies and reveals the concerns from the perspective of users, to complement those summarized from legal documents. We then track other publications that cite it. This has yielded 13 publications on the impact of GDPR on privacy policy study [11, 12, 51, 53, 70, 74], privacy policies corpus [6, 61, 71] and privacy policy language modeling [34, 37, 45, 50].

2.1.2 Quality concerns. We review these collected materials, and collect the cases that are considered as *quality concerns* (QCs), as detailed below.

QC1. Noncompliance of disclosure. GDPR Article 13 [41] and Article 29 Working party [28] clearly state the extent of information to be disclosed to the data subject. Two studies [48, 51] particularly focus on the policies' contents in terms of the disclosure of personal data collection. They identify several issues of noncompliance between the disclosed information and regulations. For example, users are not fully notified about what information is collected by the application [54]. Therefore, we identify the *noncompliance of disclosure* as one of the quality concerns that we aim to investigate in this work.

QC2. Out-of-date information. Many skills adopt a short release cycle to enable faster response to the market changes and customer needs [19], and their data handling behaviors may be altered often. California Consumer Privacy Act (CCPA) [59] requires that a privacy policy should be updated timely (at least once every 12 months) to reflect the company's most recent practices. Besides the CCPA requirement, a user study [43] reveals that whether the privacy policy is kept up-to-date concerns the public as well.

QC3. Inaccessible resources. Many skills resort to external sources in their privacy policies to provide auxiliary information. For example, some policies often include links to a third-party website, to direct users to the entity with whom the personal information is shared. However, they may miss updating these links when the external party has disabled the website link, such that the users may lose relevant information. Several studies [54, 58] have revealed that such failures are not uncommon and have raised users' concerns. Therefore, the accessibility of links in the privacy policies is identified as a quality concern in our work.

Another resource we take into consideration in this work is a skill's multi-language versions. According to Article 29 Working Party [28], "*a translation in one or more other languages should be provided where the controller targets data subjects speaking those languages*". The skills that declare a list of supported languages in their descriptions are thus supposed to also provide a set of privacy policies in these languages.

QC4. Obscure, complex, and lengthy texts. Several studies express users' difficulty with the language of the privacy policies [56, 57, 73]. Features that mostly affect readers' comprehension include grammatical and syntactic features, which refer to words and sentence structures in the privacy policies, and organizational features, which refer to document characteristics such as document length and logical order of information presentation [55, 57].

2.2 The Taxonomy

With the identified quality concerns, we generalize each of them into a quality assessment taxonomy. This results in four high-level metrics, including *compliance of disclosure* (mapping to QC1), *timeliness* (QC2), *availability* (QC3), and *readability* (QC4). We further break down each of them into items that could be measured with automatic techniques.

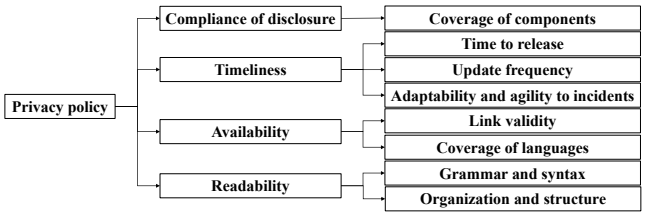


Figure 1: A taxonomy of privacy policy quality metrics.

2.2.1 Compliance of Disclosure. The compliance of disclosure of a privacy policy document refers to the extent that its disclosed information fulfills the requirement of regulations, store operators, and users. In this regard, GDPR Article 13 [40] presents the contents that are required to be included in a privacy policy. Several recent studies [11, 34, 51–53, 74, 81] also summarize types of contents from derived from data regulations or user studies. We summarize these types into 11 significant components (detailed in Table 1), and QuPer assesses the compliance of a privacy policy based on its coverage of these components.

Coverage of components. Several studies have attempted to identify significant privacy concerns by conducting user studies [34], or analyzing requirements of data regulations [51–53, 74, 81]. By reviewing these studies, we have identified eight essential components of a privacy policy that are their common concerns, including *Access, Choice, Collect, Cookies, Purpose, Retention, Security,* and *Share*. Second, we turn our consideration to the spectrum of VPA apps' user groups. Prajapati et al. [60] reveal that the lower uptake of smartphones among children and the elderly can be primarily attributed to intricate Human-Computer Interaction (HCI) models, rather than limitations in their cognitive abilities. Given that VPA apps boast a conversational UI that inherently demands minimal learning aptitudes and behavioral comprehension, they tend to attract a higher proportion of children users than their mobile counterparts, as shown in a recent work [42]. We thus include *Children* [70] as a required component. Third, as skills are open to a wide range of regions (some of which are under the protection of specific regulations, e.g., CCPA for US California users), *Region* [21] is included. Fourth, as skills are typically updated relatively frequently, we include *Update*. Finally, *Provider* is included, as the Alexa store obligates developers to disclose their contacts, to enable users to request deleting collected information.

2.2.2 Timeliness. For the timeliness, three aspects are assessment.

Time to release. Once an app is released, the associated privacy policy is supposed to be available. Therefore, we assess the difference between the first release time of the skill and that of its privacy policy.

Update frequency. The skill has undertaken many updates and some updates may introduce changes in its data handling practices. Accordingly, its privacy policy should be updated to reflect the change. We thus monitor the whole life cycle of the skill, and check whether its privacy policy is kept up-to-date.

Table 1: List of the main components that a skill privacy policy should cover

	Component	Description	Example
1	COLLECT	Types of personal data collected by the skill	We may collect: your name, birth date, gender, email address, zip code, and any other information you may voluntarily provide to us.
2	COOKIE	Cookie from user's device collected by the skill	Some service providers use cookies or similar tracking technologies in order to provide you with promotions or other contents on the basis of your browser activities and interests.
3	SHARE	How the skill shares user information	In some cases, your personal data is shared with third parties.
4	SECURITY	How the skill protects user information	These security measures include: password protected directories and databases to safeguard your information, SSL (Secure Sockets Layered) technology to ensure that your information is fully encrypted
5	RIGHT	Users' rights to their own data	You have the right to stop the advertising messages that we send to you at any time.
6	CHILDREN	Privacy Policy for Child Protection Mechanisms	We are in compliance with the requirements of COPPA (Children's Online Privacy Protection Act), we do not collect any information from anyone under 13 years of age.
7	REGION	Protection Mechanisms for Some Special Regions	If you are a resident of the state of California, we will abide by the regulations of CalOPPA when handling your information.
8	UPDATE	Whether the privacy policy will be updated	Please note that this Privacy Policy may be periodically updated. Please refer to our website for the latest Privacy Policy that is in force.
9	PROVIDER	Contact information of the privacy policy provider	If you have questions of your personal data, you may raise them at any time by contacting us at: xxx@gmail.com.
10	RETENTION	How long will the skill keep user data	We will retain user-provided data for as long as you use Bathroom Sidekick and for a reasonable time thereafter. We will retain Automatically Collected information for up to 24 months and thereafter may store it in aggregate.
11	DATA_USE	How the skill will use user data	The information we collect is used to improve our website in order to better serve you.

Adaptability and agility to incidents. Some events may stimulate developers to update the privacy policies. Security incidents, e.g., a data breach, may also raise alerts to developers for the update. We thus analyze the relevance of the releases and updates on privacy policies with the occurrence of known security incidents.

2.2.3 Availability. We consider the availability of the following two resources according to the identified quality concerns.
Link validity. This refers to the accessibility of the external links.
Coverage of languages. This refers to whether a skill provides privacy policy in the language versions as it claims in the list of supported languages.

2.2.4 Readability. As identified in the quality concerns, we aim to assess a privacy policy's readability from the two features that mostly affect users' comprehension.
Grammar and Syntax. This focuses on the effect of textual elements (e.g., words and sentences) on comprehension. We include most of the representative features discussed in relevant studies [28, 57].

Avoiding double negatives. A double negative means that a positive statement uses two negative elements to produce a positive force. For example, "*We will not share your information with organizations or institutions that we do not work with*" actually means "*We will share your information with organizations or institutions if we work with them.*" According to a relevant study [57], the latter one is easier to comprehend.

Avoiding obscure language qualifiers. Article 29 [28] requires that "*Language qualifiers such as 'may', 'might', 'some', 'often' and 'possible' should also be avoided*". For example, the statement "*We may use your personal data for research purposes*" violates this requirement.

Locating the main idea of the sentence at the beginning. Sentences are easier to read and comprehend when the main idea occurs at the beginning [57]. For example, "*We do not share personal information such as name, address, email address, or phone number with others*" can be better understood than "*Personal information such as name, address, email address, or phone number is not shared with third parties by us*".
Organization and Structure. This focuses on the document features, such as the length and logical order of the information.

Text structure. Structural features that may affect readability include the number of sentences in a paragraph, the number of words in a sentence, and the number of syllables in a word, as revealed by [29, 67]. These features represent the difficulty level of the paragraph, the sentence, and the word, respectively.

Logical order of information. The content should be arranged in a way that is presented in a logical order [57]. For example, the COLLECT section in Table 1, which builds the context of what information is being collected, should appear in front of the SHARE section, which describes how the skill shares user information, to facilitate readers' understanding.

3 OVERVIEW OF QUPER

We embed the proposed quality metrics into a framework named QuPer, to assess the quality of privacy policies. In this section, we brief the process of privacy policy collection (Section 3.1) and the assessment techniques for each quality metric (Section 3.2).

3.1 Data collection and preprocessing

We obtain a list of all skills available in the Alexa skill store from the dataset used in a recent study [81]. It includes 65,195 skills, of which 21,063 skills provide links to their privacy policies. We filter out the skills that provide duplicate privacy policy links, and 9,136 skills are left. We use a crawler to scrape the privacy policy documents through the obtained links. During the crawling, 584 URLs cannot be opened and 1,245 URLs return the pages of 404 not found. As such, 7,307 privacy policy documents are obtained. Among the obtained documents, some are not related to privacy policies (e.g., a company home page). We thus filter out those that do not include the keyword "privacy" or "user information". To ensure the validity of the filter, we conduct a manual confirmation on 200 policy links randomly selected. The results are listed in Table 2. Overall, 6,430 privacy policies are kept and formulate the cohort for our timeliness and availability assessment.

We further build a crawler to retrieve the skill home pages and use the Beautiful Soup library [2] to extract the language version information of privacy policies from the home pages for the assessment of the availability of supported language. Among the 6,430 privacy policies, we further filter out the privacy policies that are not presented in HTML format and leave 5,473 documents for the completeness and readability assessment (as QuPer relies on the HTML structure, e.g., HTML tags, to automate the analysis, as detailed in Section 4 and 7).

Table 2: Benchmarking policy collection and preprocessing

	Cases		Confusion matrix	Value	Metrics	Value
	Positives †	Negatives	False Negative (FN)	0	FP Rate	8.00%
GroundTruth	175	25	False Positive (FP)	2‡	Precision	98.87%
QuPer	177	23	True Negative (TN)	23	Recall	100.00%
			Ture Positive (TP)	175		

† Positives: legitimate policies. Negatives: web pages with content irrelevant to privacy policies.
‡ In both FP cases, the privacy policy link directs to the company's homepage, where the menu bar contains the keywords "privacy" or "user information".

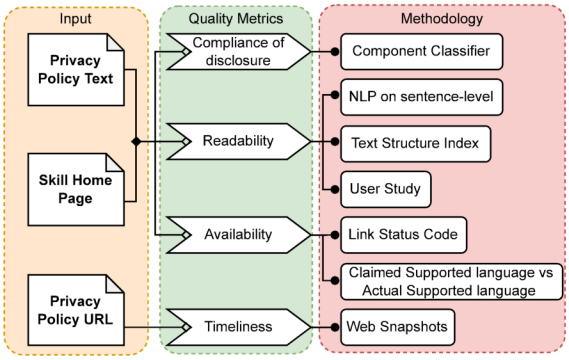


Figure 2: Overview of QuPer

3.2 Design of QuPer

Figure 2 shows an overview of QuPer's methodology. Based on the preprocessed documents, QuPer extracts three types of information from the privacy policies, i.e., (1) privacy policy texts which are used for the analysis of disclosure compliance, readability, and availability (valid link) analysis, (2) privacy policy URLs which are used to analyze the timeliness of privacy policy, and (3) skill home page for availability (supported languages) analysis. They are taken as the input for the assessment, and we propose the following techniques for each of the quality metrics.

Disclosure compliance assessment. QuPer's compliance analysis has to understand the meaning of privacy policies, and thus it relies on identifying the semantics of privacy policy texts. Nevertheless, those privacy policies are written in natural languages by various developers, and turn out to be heterogeneous in their patterns, wordings, and formats (***Challenge #1***). Naive methods that are based on keyword matching often lose the context information. Existing work on the privacy policies of web pages and mobile applications [11, 12, 53, 72, 76] may also fall short in providing a comprehensive assessment in the VPA context. A typical challenge is that some quality concerns we target to analyze, such as children- and region-specific policies, are not covered by them, such that they cannot be straightly applied. We thus propose a lightweight hybrid method to precisely identify the components covered by the privacy policy texts. It consists of a section-level analysis based on machine learning techniques to capture the context, and a sentence-level analysis based on NLP techniques to leverage sentence structure. The compliance analysis is detailed in Section 4.

Timeliness assessment. In the domain of mobile apps, explicit disclosure is required for each update [33], such as specifying added and removed features and updating data handling processes. In contrast, the skill store currently operates without equivalent regulations. To analyze the timeliness of a privacy policy, QuPer has to know the update history of both skills and their privacy policies. Such information is not made publicly available by the Alexa store though (***Challenge #2***). To address this challenge, QuPer leverages the digital Internet archive provided by Wayback Machine [1]. Wayback Machine keeps track of web page snapshots that are periodically[1] taken over the last 20 years, and the archived web pages even include those that no longer exist. From the archives, QuPer obtains the release time of skills, and the update dates of their privacy policies, covering the time period from 2018 (when Alexa skill service became public for testing) to 2022. The timeliness analysis is detailed in Section 5.

Availability assessment. When assessing the accessibility of the external links in the privacy policy, QuPer may detect irrelevant links (e.g., a link in the displayed advertisement) from the web page (***Challenge #3***). QuPer filters out them based on the context information, by querying our section-level classifier (see the completeness analysis) on the correlation of the sentence that the link is located in.

For the availability of supported language, the challenge is to recognize all language versions of the privacy policies that the skill releases (***Challenge #4***). QuPer embeds multiple detection mechanisms to address this. It checks all links in the "*supported language*" section on the skill's home page of the skill, detects the functionality of language switching, and mutates the language tags in the URL (e.g., *en* to *de*). The availability analysis is detailed in Section 6.

Readability assessment. VPA is designed to enhance accessibility and inclusivity, so its users include children and individuals with low privacy awareness and limited ability to understand complex privacy documents. This highlights the importance of ensuring that VPA privacy policies are easily understandable and acceptable to users across most age groups and expertise levels, emphasizing the need for high readability.

The readability of a document can vary depending on its characteristics, e.g., the complexity of the topic, and a few factors of readers, e.g., expertise and reading ability. Evaluating readability thus can be subjective (***Challenge #5***). We turn to the literature on the readability of general documents and volunteers of IT and legal background. To evaluate the structural features, we adopt three commonly used metrics such as Automated Readability Index (ARI) [67] (aiming to assess the required grade level of education to understand a document), and seven metrics that well represent the structure of a document such as Letters Per Word (LPW), Syllables Per Word (SPW), and Words Per Sentence (WPS). We also consult volunteers with relevant expertise based on a classic study on readability [30] to understand the users' preference on the logical order of sections, and compare the actual order of presentation in the privacy policy under assessed. The readability analysis is detailed in Section 7.

4 DISCLOSURE COMPLIANCE ASSESSMENT

In this section, we present our method for assessing the compliance of disclosure (Section 4.1) and our findings regarding this quality metric (Section 4.2).

[1]The time period between two snapshots is not on a regular basis, as per Wayback Machine has clarified in the response to our queries.

Table 3: HTML heading tags captured in privacy policies

Subheading tag	h1	h2	h3	h4	h5	b	strong
Percentage	12%	26%	18%	8%	7%	11%	18%

4.1 Compliance Assessment Methods

To address the challenge posed by the unformatted and heterogeneous nature of privacy policies (***Challenge #1***), we design our assessment methods by combining machine learning and NLP techniques. Considering that our dataset is relatively small-scale, we propose a lightweight and fine-grained approach so that it can be more precise than using a pure multi-label multi-class classification. Our approach involves analysis at both the section and sentence levels, aiming to capture precise information about the sentences in the privacy policies.

4.1.1 Section-level analysis. The privacy policy is usually structured into sections, and each section covers one of the components listed in Table 1. For example, the component of *COLLECT* is often presented in a section that is titled semantically similar to "*Information we collect*". The vast majority of privacy policy documents (>90%) follow this format [81]. Therefore, we propose to use the section titles as input to train a classifier that categorizes the purpose of the section (according to Table 1) to assess whether the document includes the corresponding components.

Section Title Extraction. Section titles are typically represented as heading elements in HTML and can be identified by their surrounding tags (listed in Table 3). These tags cannot be used as filters straightly though, as some irrelevant contents, such as advertisements and navigation bars, may also use HTML tags to highlight the text. We thus need to identify the particular type tag which is used to highlight the section titles in the policy texts. To this end, we first define five reference phrases, which are the top five most frequently appearing relevant phrases in privacy policy section titles (*"information collect"*,*"information use"*, *"change data"*, *"data security"*, *"contact us"*). We then extract the text embedded in heading tags from the privacy policy in HTML format and use Jaccard similarity coefficient [36] between the extracted text and reference phrases to identify section titles, as defined below.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

where A denotes an extracted phrase and B denotes a reference phrase. For example, in a privacy policy that contains both "h2" tags and "h3" tags, texts extracted from "h2" tags are "News", and "Blog", both of which are of zero Jaccard score to the reference phrases. Texts extracted from "h3" tags in the document are "What information we collect", "Security of data", and "Cookies", which are of 0.50, 0.66, and 0.0 Jaccard scores to the reference phrases, respectively. We treat the tag with the highest average value of the Jaccard score as the section title tag. That means, in the above example, we identify the "h3" tag as the section title tag. We then extract the texts which are highlighted by the identified tag as section titles.

Section Title Classification. With the extracted data, we build a classifier to classify privacy policy sections into the components defined in Table 1.

Table 4: The Precision (P), Recall (R), and F1-score (F) of section-level classifiers (%)

	Naive Bayes			SVM		
Component	***P***	***R***	***F***	***P***	***R***	***F***
COLLECT	91.67	68.75	78.57	90.71	69.23	81.82
COOKIE	83.68	98.89	84.85	73.68	81.97	84.85
SHARE	85.71	90.91	88.24	82.86	80.63	76.57
SECURITY	82.14	92.00	86.79	82.14	92.00	86.79
RIGHT	77.78	77.78	77.78	72.22	86.67	78.79
CHILDREN	80.43	84.95	80.00	90.48	83.73	85.53
REGION	79.97	85.71	70.59	80.53	84.69	79.74
UPDATE	80.37	90.91	85.11	72.00	83.91	73.72
PROVIDER	95.45	84.00	89.36	95.45	87.50	81.30
RETENTION	85.71	98.69	92.31	64.29	91.37	72.26
DATA_USE	83.71	72.73	76.19	86.67	81.25	83.87
OTHER	76.92	67.80	72.07	80.77	64.62	71.79
Average	**83.63**	**84.84**	**81.82**	80.98	82.30	79.75

Table 5: Most frequently used predicative verbs of components and average pair-wise similarity

Component	Most frequently used predicative verbs	Average pair-wise similarity
COLLECT	collect, gather	0.68
COOKIE	cookie	1.00
SHARE	share, disclose, sell, trade, transfer, send	0.39
SECURITY	guard, protect, guarantee, defend, preserve, safeguard	0.52
RIGHT	give, empower, grant	0.34
CHILDREN	address, aim, intend	0.43
REGION	locate, live, settle	0.33
UPDATE	change, update	0.45
PROVIDER	send, contact	0.40
RETENTION	store, retain, maintain	0.44
DATA_USE	use, utilize, operate, exercise, practice, apply	0.52

Data labeling. We invite five researchers from our institution to conduct the data labeling. All of them have research experience in privacy policies and one has a law background. To ensure the accuracy of data annotation, we first provide them with a brief tutorial and some annotation samples. We ask them to label 48 section titles in five privacy policies and explain their labeling in a group discussion, ensuring that their criteria are calibrated. After that, we randomly select 140 privacy policies and ask them to annotate all included (1,503) section titles with 12 labels (i.e., the 11 components given in Table 1, plus the label "OTHER").

Training. We use the support vector machine (SVM) [38] and the Naive Bayes classification to train our classifiers, given that both are known to have strong capacity in handling relevant tasks [17, 65]. The performance of the classifiers is shown in Table 4, in which the Naive Bayes classifier (81.82% F1-score) achieves a higher accuracy on average compared to SVM (79.75% F1-score).

4.1.2 Sentence-level analysis. The section-level analysis can infer the purpose of each section, and we conduct sentence-level processing as complementary. This is to handle the obstacle that a section contains multiple components or two section titles overlap. For example, a section named "Collection and use of information" contains both *COLLECT* and *DATA_USE* contents ("*We collect your information such as name, email, and address. We use such information only for statistical purposes that help us design and administer the Site*"), which causes the section-level classification to have a low recall on the *COLLECT* component, as shown in Table 4.

Our sentence-level analysis begins with defining the categorization thresholds. To this end, we adopt Spacy [3] to retrieve the most frequently used predicative verbs for each component as shown

in the second column of Table 5. We then calculate the average pair-wise similarity of the verbs as shown in the third column, and set those similarities as thresholds to categorize sentences from policy documents.

Next, we extract sentences that contain a predicative verb and keywords "information" and/or "data" from a privacy policy. We then calculate the similarity between the predicative verb in the extracted sentence and the most frequently used predicative verb in each component (column 2 in Table 5). We assign the sentence to a component if the similarity score is higher than the threshold.

To examine the performance of our compliance assessment approach before applying it to a large-scale study, we randomly select 100 privacy policies and manually confirm QuPer's classification. As the results listed in Table 6 show, QuPer achieves high performance in both precision and recall.

4.2 Compliance Assessment Results

We first investigate the component coverage among 5,473 privacy policies. As shown in Table 7, 63% of the privacy policies cover the *COLLECT* component, while less attention is paid to *RETENTION* (16%) and *REGION* (20%) related policies. In addition, only 28% of privacy policies provide *CHILDREN* related content, which raises an alert to the app developers and users (parents) in terms of the children's privacy. Similar concerns have also been raised by recent studies [18, 82] that kids can access and spend a long time on VPA apps and they are also more vulnerable to threats compared to adults. In addition, we highlight that only 1.17% privacy policies provide complete categories in Table 1.

When a privacy policy has a *COLLECT* component, we consider that the corresponding skill collects information from the user. Article 13 of GDPR requires that if a data controller collects personal data from a data subject, it should also provide the following information to the data subject at the time of obtaining the personal data:

- *The contact details of the data protection officer → PROVIDER*
- *The purposes of the processing for which the personal data are intended → DATA_USE*
- *The controller intends to transfer personal data to a third country or international organization → SHARE*
- *Reference to the appropriate or suitable safeguards and the means by which to obtain a copy of them or where they have been made available → SECURITY*
- *The period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period → RETENTION*
- *The existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject or to object to processing as well as the right to data portability → RIGHT*

As such, we select 3,371 policies (among the total 5,473 privacy policies) which provide the *COLLECT* component for the GDPR compliance assessment. Table 8 shows the coverage results, which is concerning, as we observe that a large number of assessed policies do not comply with the regulation. Even the best coverage result only reaches 68%, indicating that a significant proportion (nearly one-third) of those privacy policies that include the *COLLECT* component do not provide information about *DATA_USE*. The coverage of other categories is even worse, with the most overlooked one as *RETENTION*, of which the coverage is as low as 22%.

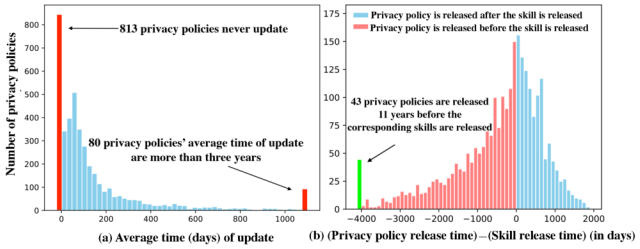


Figure 3: Statics of release and update frequency

Finding 1: We find that 90% (3,052/3,371) privacy policies of skills miss covering at least one component. Only 28% of privacy policies include content about children's privacy, which is contrary to the CCPA's requirements. The most overlooked component is *RETENTION*, as we only find 22% of privacy policies cover it although it is mandatory according to GDPR. The lack of *RETENTION* policy puts users at risk as their data may be kept indefinitely as long as the developers wish without notice.

5 TIMELINESS ASSESSMENT

In this section, we present our method for timeliness assessment (Section 5.1) and report our findings in this regard (Section 5.2).

5.1 Timeliness Assessment Methods

As the Alexa store does not release the update history of skills and their privacy policies (***Challenge #2***), we turn to Wayback Machine, an Internet archive service that tracks website updates. It periodically scrapes and achieves the snapshots of web pages. Given a URL, it lists all captured updates on the texts of web pages along the queried period.

We first obtain snapshots of all pages of the privacy policy from the time it was posted for the first time until the last update, and then assess the update frequency by counting the number of unique web pages in between this time period. We exclude 594 privacy policies that are generated by dynamic pages, as they cannot be recognized by WaybackMachine, and eventually obtain the update history of 4,879 privacy policies in total for the timeliness assessment.

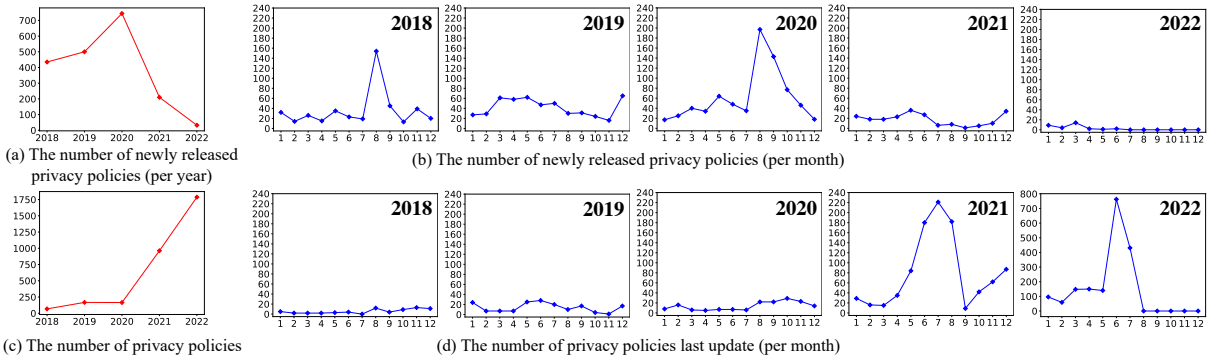
5.2 Timeliness Assessment Results

Update frequency. Figure 3(a) shows the frequency distribution of privacy policy updates. The X-axis represents the average time of privacy policies' update (in days) and the Y-axis represents the number of privacy policies. We highlight that 813 privacy policies (16%) have never been updated (167 since the year 2022, 192 since 2021, 138 since 2020, 81 since 2019, and 235 since 2018 and earlier), and 89 privacy policies' average time of update are more than three years. We also observe that only 41% of privacy policies are updated within 100 days. 14% of privacy policies are updated between 100 to 200 days, while 29% of privacy policies are updated more than 400 days.

Table 6: QuPer's performance in identifying required components in privacy policies

	COLLECT		COOKIE		SHARE		SECURITY		RIGHT		CHILDREN		REGION		UPDATE		PROVIDER		RETENTION		DATA_USE	
	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG	POS	NEG
Groundtruth	63	37	31	69	44	56	45	55	37	63	36	64	19	81	39	61	42	58	15	85	41	59
QuPer	61	39	32	68	48	52	49	51	38	62	39	61	20	80	42	58	44	56	15	85	41	59
FP rate	2.70%		2.90%		8.93%		9.10%		1.59%		6.25%		2.47%		4.92%		6.90%		1.18%		3.39%	
Precision	98.36%		93.75%		89.58%		89.80%		97.37%		89.74%		90.00%		92.86%		90.91%		93.33%		95.12%	
Recall	95.24%		96.77%		97.73%		97.78%		100.00%		97.22%		94.74%		100.00%		95.24%		93.33%		95.12%	

† POS stands for the number of positives and NEG stands for that of negatives.



(a) The number of newly released privacy policies (per year)

(b) The number of newly released privacy policies (per month)

(c) The number of privacy policies last update (per year)

(d) The number of privacy policies last update (per month)

Figure 4: Trend of privacy policies' release/update from 2018 to 2022

Table 7: Component coverage among 5,473 privacy policies

Component	Frequency & Coverage	Component	Frequency & Coverage
COLLECT	3371 (63%)	REGION	1085 (20%)
COOKIE	2435 (46%)	UPDATE	2258 (42%)
SHARE	3215 (60%)	PROVIDER	2061 (38%)
SECURITY	2724 (51%)	RETENTION	859 (16%)
RIGHT	2144 (40%)	DATA_USE	2906 (54%)
CHILDREN	1484 (28%)		

Table 8: Component coverage of the privacy policies which include the COLLECT component

Component combination	Frequency	Coverage
COLLECT & PROVIDER	1826	54%
COLLECT & DATA_USE	2304	68%
COLLECT & SHARE	1717	51%
COLLECT & SECURITY	2132	63%
COLLECT & RETENTION	736	22%
COLLECT & RIGHT	1642	49%

Privacy policy release time vs. Skill release time. Figure 3(b) shows a comparison of privacy policies' release time and their corresponding skills' release time. We find that more than half of the privacy policies are published before the corresponding skills are released, among which 43 privacy policies are released 11 years earlier than the skills. This occurs mostly because the developer straightly directs their users to the privacy policy of their other services (e.g., the website), rather than specifically creating a precise one based on the data handling behaviors of the skill. This should raise an alert to the public because Alexa skills usually collect user information in a broader range of ways (such as recording the user's voice) than traditional web services.

In addition, we find that a number of privacy policies are released long after their skill are published, e.g., more than one hundred privacy policies are released three years after the release of their corresponding skills. During this period, the skill's behavior is not governed by the privacy policy, which puts users' privacy at risk.

The trend of privacy policies' release and update. We investigate the trend of privacy policies' release and update from 2018 to 2022. Figure 4(a) shows that the number of newly released privacy policies peaks in 2020, followed by a rapid decline in 2021 and 2022. This result is consistent with the skill releases trend as shown in Voicebot report [44]. Although the number of skills is still increasing, its growth rate has decreased since the end of 2019.

Figure 4(c) shows that most privacy policies' updates took place in 2021 and 2022, while 197, 186, and 78 privacy policies have their last update time in 2020, 2019, and 2018, respectively.

Adaptability and agility to incidents. Figure 4(b) and 4(d) demonstrate the number of policy releases and updates per month from 2018 to 2022. As shown in Figure 4(b)-2018, the first peak of policy release is in August 2018. This can be because Amazon allows skill providers to add *in-skill purchases* and *Amazon Pay* functions in Alexa skills in May 2018 [31], which stimulates the skills' profit growth and results in an influx of skill providers. Another peak of policy release appears in September 2020. This aligns with the rapid growth of skill users, which peaks in early 2020 [44].

In Figure 4(d), we observe that the first peak of policy updates occurs in July 2021. This is the time when Luxembourg National Commission for Data Protection (CNDP) levied the largest GDPR violation fine of 746 million Euros against Amazon [20]. We also find that, in 2022, the number of updated privacy policies is relatively high from January to August, since more than 30 incidents of data leakage occurred during that period [22].

Finding 2: We observe that 16% (813/4,879) of privacy policies are never updated. We find that 813 Amazon skills have been released for more than two years and have never updated their privacy policies.

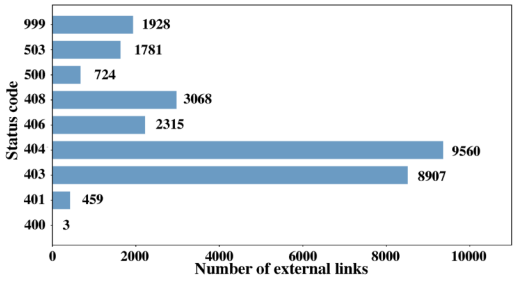


Figure 5: Status code distribution of abnormal sites

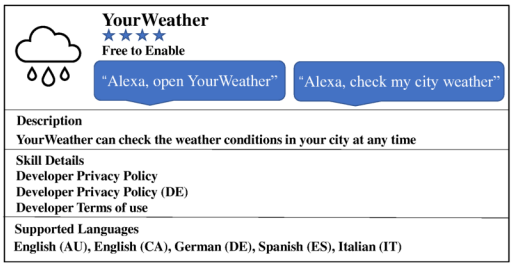


Figure 6: Supported languages on skill homepage (example)

Finding 3: The release of a privacy policy does not align with the release of its skill. We find that 449 skills directly reuse the privacy policies from their developers' company websites which are usually published long before the release of the skill. Furthermore, 1,503 skills release their privacy policies long after the release of the skills.

Finding 4: The number of policy updates abruptly increases right after the occurrence of security incidents. Many skill developers update their privacy policies only after data security incidents occur, which fails to comply with the regulation's requirement of timely updates to reflect the company's current practices [59].

6 AVAILABILITY ASSESSMENT

In this section, we present our method of availability assessment and our assessment results.

6.1 Link Validity

Method. Privacy policies can provide external links in two ways: (1) directly display the in-text link (i.e., "*For more information, you can visit our website www.xxx.com*"), or (2) use HTML "href" tags (i.e., "*Please click here for more information*"). For the former, we extract the external links through pattern matching with regular expressions. For the latter, we retrieve the links by extracting all "*href*" elements from the HTML code of the webpages. We then implement a script enabled by a Python library named "*urlopen*" to automatically check the links' availability. The availability of a link is assessed by getting the status codes (detailed in Table 10 in Appendix A) returned from the websites.

Results and Findings. We have obtained 200,183 external links in total from 6,430 privacy policies. Our availability assessment finds that 171,438 (85.7%) external links can be visited successfully (i.e., status code 200 returned), and 28,745 (14.3%) external links are not accessible. We present the distribution of non-accessible links by the returned status codes in Figure 5.

We observe that the most common status code returned by non-accessible external websites is 404 (i.e., the *webpage not found* error). In addition, more than 8,000 links return the 403 status code (i.e., the *forbidden* error), which means the website resources are not publicly available.

Finding 5: We find that 14.3% (28,745/200,183) of external links extracted from Alexa skills' privacy policies are invalid. Among 6,430 privacy policies, 3,788 (58.91%) contain at least one invalid external link. We also find that among the privacy policies containing invalid external links, 55 (1.15%) have more than 100 invalid links.

6.2 Supported Language Checking

Method. The supported language checking aims to assess if the actual supported languages in the privacy policy sufficiently cover the supported languages declared by the skill. The list of declared supported languages can be retrieved from the skills' home pages (as shown in Figure 6). However, there does not exist a canonical way for skill owners to provide the list of actual supported languages of privacy policies. For the actual supported languages, we find that skill owners usually release their privacy policies in different languages in two ways: (1) directly providing a link on the home page, or (2) providing a language selection button in the web page to redirect users to the other language versions. We randomly select 50 privacy policies that include versions in other languages to check the effectiveness of the two methods, and find that both can retrieve all language versions of the privacy policies. We first attempt to directly retrieve the links from the HTML source code (see Section 5.1). Thus, we can obtain the language list if it is directly given on the web page. In case the language list is provided in the latter manner, we will first locate the language-selection buttons on the web page by matching keywords, such as "English" and "Language". We then simulate button-clicking events using the Python Selenium library [5] to find whether other language versions of privacy policies are available. Since some English privacy policies URLs contain the "en" keyword, we replace them with "fr" (French), "de" (German), and other language abbreviations to test if the modified URLs can be accessed and display the corresponding versions of the privacy policies.

We use $\mathcal{A}$ to represent the set of actual supported languages, and $\mathcal{B}$ to denote the declared supported languages. If $\mathcal{B}$ is a subset of $\mathcal{A}$, we consider that the skill developer covers all the language versions of privacy policies that are declared in the skill's "supported language" section. Otherwise, we consider that the skill developer does not provide the required language versions for privacy policies.

Results and Findings. We find 2,602 (40%) privacy policies do not provide the language versions as declared in the skills' "supported language" sections. Among them, 1,667 skills are missing one language version of the privacy policy, and 752 skills are missing two languages. 183 skills are missing three or more languages of the privacy policy. Recent statistics [77] show that the Spain skill store and Germany skill store have the top skill counts among all the Alexa skill stores in non-English speaking areas. We also find that

the largest number of missing language versions of the privacy policy is in Spanish, followed by German.

Finding 6: We observe that 40% (2,602/6,430) skills do not provide required privacy policy language versions as they declared in the "Supported Language" sections.

7 READABILITY ASSESSMENT

In this section, we present our readability assessment methods (Section 7.1) and our results (Section 7.2).

7.1 Readability Assessment Methods

We examine the grammatical and syntactic features including *double negative*, *obscure language qualifiers*, and *main idea's location*, and the organizational features including *sentence length* and *logic order of information*. We now detail them respectively.

Avoiding double negatives. It occurs when a sentence contains two grammatical negation forms. Therefore, to detect double negative sentences, we use keyword matching to check the number of negative words and contrast words in a sentence. We use the negative and contrast vocabulary word list as the reference for the identification [25, 26, 39].

Avoiding obscure language qualifiers. The Article 29 Data Protection Working Party [28] states that "language qualifiers such as "may", "might", "some", "often" and "possible" should be avoided. Therefore, we use these keywords to conduct fuzzy semantic tests on the sentences of privacy policies.

Locating the main idea of the sentence at the beginning. We determine the main idea of a sentence by locating the subject and predicate in this sentence [49]. We resort to the Python SpaCy library to find the index position of the subject (nsubj) and predicate (ROOT) in each sentence, and then use the following Equation 2 and 3 to calculate the positions of the main idea (I_c) depending on the length of the sentence.

$$I_c = \frac{I_s + I_p}{2}, \tag{2}$$

where I_c is the central index, I_s is the index of subject, and I_p is the index of predicate. We add the index of the subject and the predicate and then divide it by two to get the central index.

$$M_{True} = 1, \; if \begin{cases} I_c < L_s/2 & : 20 > L_s > 5 \\ I_c < L_s/3 & : 27 > L_s >= 20 \\ I_c < L_s/4 & : L_s >= 27 \end{cases} \tag{3}$$

where M_{True} is a Boolean variable used to determine whether a main idea is at the beginning.

We do not include the cases in which the sentence length is less than 5 as they are rare according to the literature [69]. When the sentence length is greater than 5 and less than 20, we compare the central index I_c with half the sentence length. When the sentence length is greater than 20 and less than 27, we compare the central index to one-third of the sentence length, and we compare the central index to a quarter of the sentence length when the sentence length is greater than 27. We apply this method to 20 randomly selected sentences and achieve 90% accuracy in identifying the main idea location. Table 9 shows two examples in which the second sentence achieves a lower central index I_c than the first one.

Table 9: Example of main idea sentences

Sentence 1	Token	Token POS	Sentence 2	Token	Token POS
with	prep	ADP	we	nsubj	PRON
your	poss	PRON	share	ROOT	VERB
permission	pobj	NOUN	your	poss	PRON
,	punct	PUNCT	email	compound	NOUN
we	nsubj	PRON	address	dobj	NOUN
share	ROOT	VERB	with	prep	ADP
your	poss	PRON	your	poss	PRON
email	compound	NOUN	permission	pobj	NOUN
address	dobj	NOUN			
central index	5.5		**central index**	1.5	

Text structure. We first adopt three readability metrics that have been widely used by other studies [29, 64] for measuring the document readability, namely Automated Readability Index (ARI), Flesch Readability Ease Score (FRES), and Laesbarheds Index (LIX). The first metric, i.e., ARI, is used to assess the required reader's education level to understand a document, and the other two, i.e., FRES and LIX, assess the difficulty level of the document based on its average number of syllables per word, total words, and total sentences. In addition, we select seven metrics from existing studies [15, 29, 66, 68, 78] that specifically focus on assessing the impact of document structure on readability. These metrics have been revealed to have a significant impact on readability, and include letters per word (LPW), syllables per word (SPW), words per sentence (WPS), sentence count (SC), word count (WC), reading time (RT) and speaking time (ST). In Table 11 of Appendix B, we list details of all ten metrics QuPer considers.

Logical order of information. Recent studies show that a coherent organization of sections/paragraphs in a document can largely ease readers' comprehension [7, 30, 75]. From existing literature, we summarize the following guidelines of the presentation order.

- *Put content in a time sequence.*
- *Present the general information before the specific one.*
- *Discuss things that affect many people before those that affect few.*
- *Present permanent provisions before temporary ones.*

Based on these guidelines, we then investigate a desirable arrangement of the section in a privacy policy through a user study. We recruit 23 volunteers, ten of whom have a major in law and have experience in writing legal documents. The other 13 major in computer science. All of them have experience in reading privacy policies. We prepare a tutorial with the summarized guidelines for the volunteers, and ask them to sort the sections in the order they find logical for them to interpret the policy. We use the majority vote among the volunteers to produce the final order of the sections. Our study has been guided by an ethics committee member in our university. We list the tutorial in Table 12 in Appendix C.

7.2 Readability Assessment Results

Results of grammatical and syntactic issues. Figure 8 shows the distribution of privacy policies including double negative sentences (3,086), privacy policies including obscure language qualifiers sentences (4,687), and privacy policies including sentences whose main ideas are not at the beginning (4,139) among 5,473 privacy policies.

- **Avoiding double negative.** We observe that 56% (3,086/5,473) of skills' privacy policies contain double negative forms. Among the 3,086 privacy policies, 2,010 privacy policies contain no

Figure 7: Actual section order vs. user's desirable order

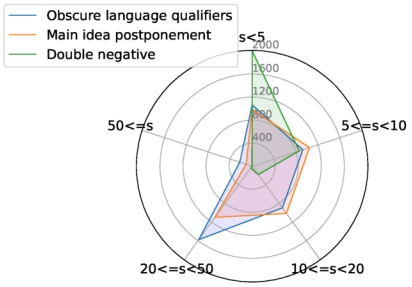


Figure 8: Results of "Double negative", "Obscure language qualifiers" and "Main ideas are not at the beginning"

more than 5 double negative sentences, and 807 privacy policies contain more than 5 but no more than 10 double negative sentences.

- **Avoiding obscure language qualifiers** We find that 1,594 privacy policies contain more than 20 sentences that use obscure language qualifiers, which indicates that 34% (1,594/4,687) of privacy policies use a lot of fuzzy sentences to confuse users. We randomly select 10 privacy policies that contain obscure language qualifiers, and find that they commonly appear in *COLLECT* section. For example, "*We may collect your data including...*".
- **Locating the main idea of the sentence at the beginning** Sentences in which the main idea is not placed at the beginning are evenly distributed (exclude range that is over 50). We find that 76% (4,139/5,473) of privacy policies contain at least one sentence that its main idea is not located at the beginning.

Result of text structure. In Figure 9, the average value of the Automated Readability Index (ARI) is around 14, meaning that the privacy policy requires readers of a high school or even college degree to be able to understand the document (Scope mapping is in Table 11). It also shows that there are a lot of long and difficult sentences and words in the current privacy policy. For example, long sentences (WPS>27) appear frequently in privacy policies (36%), which hinders readers' understanding according to [69]. In addition, some privacy policies have a much larger number of words that are unacceptable (WC>25,000 words). We manually go through these privacy policies and find that they are full of advertisements, third-party platform information, and non-English language, among other things. These privacy policies also lead to anomalies in the FRES, LIX, and ARI values (FRES values below 0, LIX values above 1300, and ARI values above 30).

Results of logical of information. Figure 7 (top) summarizes the result of the actual privacy policy section order, in which the weight represents the average value of the ordering scores ranging from 1 to 11. For example, if a component (e.g., *COLLECT*) appears at the beginning of a privacy policy, we give it the highest ordering score (i.e., 11.). Figure 7 (bottom) presents the results of a desirable logical order from the user study.

We find that most privacy policies often put the *COLLECT* at the beginning. This is consistent with the users' expectations. After *COLLECT*, many privacy policies immediately present *DATA_USE*, of which the weight (5.81) is very similar to the *COLLECT* (5.89). However, users prefer to put *COOKIE* in the second order, followed by *SHARE*. In addition, users placed *UPDATE* and *PROVIDER* at the end, while in actual presentation *CHILDREN* and *REGION* are usually stated at the end.

Finding 7: Most sentences in the privacy policy do not conform to the user's reading habits. We find that more than half of privacy policies contain "Double negative" (56%), "Obscure language qualifiers" (34%), and "Main ideas postponement" (76%). The readability metric also shows that the current privacy policy still contains long sentences and difficult words, which requires users with a high school or college degree to fully understand its content.

Finding 8: The organization of section order generally meets the users' expectation, while *CHILDREN* and *REGION* are expected by the users to appear earlier and *DATA_USE* can appear late.

8 IMPLICATIONS AND LIMITATIONS

8.1 Implications

Our findings highlight that the quality of privacy policies among current VPAs (i.e., Amazon Alexa) is concerning. It should raise an alert to app developers, store operators, and users in VPA ecosystems. In this section, we discuss the implications of our work to these three roles, with a focus on the developers who are the primary responsible party for developing high-quality privacy policies.

Users. A privacy policy is the default agreement between the user and the developer, and it may serve as a reference upon the unfortunate event of a data breach. Therefore, users should raise awareness of the privacy policy quality, as an ill-crafted privacy policy may indicate an unfair position. Users are particularly recommended to examine the compliance of disclosure to ensure that all fundamental components are covered in the policy. For example, most skills do not disclose their practices about *RETENTION* (see Section 4.2). Users should also check the privacy policy to understand the app's data collection and management before starting to use the service.

Store operators. One main reason for the quality issues is that the app store does not enforce a strict requirement on the quality of the privacy policy documents. We recommend the operator include a quality check in the app vetting process, ensuring privacy policies fulfill the four quality metrics before accepting the uploaded apps. In addition, the store can provide a unified privacy policy template for developers, enabling those without a legal background to create privacy policies that comply with regulations.

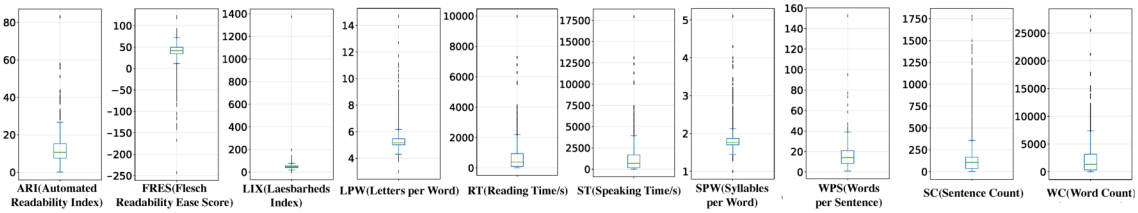


Figure 9: Result of text structure indexes

App developers. We highlight that it is desirable for app developers to craft high-quality privacy policies. Summarizing our findings, we propose the following four guidelines for developers.

Privacy policies shall cover every stage of processing personal data of data subjects. Our key findings summarized in Section 4.2 show that most privacy policies are incomplete in terms of the disclosed information, and the *RETENTION* is the most commonly overlooked component. Therefore, we recommend developers consider the entire data life cycle, encompassing not only data collection itself but also the entire data handling process.

Privacy policies shall be kept updated. The privacy policy should not be deemed as a permanent document. Instead, it should be updated with the app's evolution and alterations in laws and regulations. Nevertheless, as our findings reported in Section 5.2 and Section 6.1, some privacy policies are never updated or directly linked to irrelevant web pages. Therefore, we recommend that developers should provide a dedicated privacy policy for each of their VPA apps and keep it updated.

Developers should place emphasis on the implementation of localization strategies and practices. Providing a privacy policy version corresponding to the language supported by the app is crucial for users. Although advanced translation software is prevalent today, it is challenging to guarantee precision. Considering that privacy policies are formal documents, developers are responsible for providing precise translations. This is also a demonstration of fairness for users in different countries and regions.

Privacy policies shall provide information in logical order. In order for users to comprehend and effectively use a privacy policy, it is important to organize the presented information in a logical way. This has been largely overlooked in existing privacy policies though (see Section 7.2). Therefore, we suggest that developers should follow the order that facilitates users' reading and comprehension.

8.2 Limitations

QuPer focuses on measuring the quality of privacy policies in the context of VPA apps. To the best of our knowledge, this is the first work that conducts a systematic and large-scale analysis. However, the current work of QuPer carries several limitations that should be addressed in future work.

First, although QuPer has developed a taxonomy that takes into account a wide range of sources including major data regulations, standards, and literature, the establishment of a comprehensive taxonomy necessitates increased efforts from the research community, such as considerations for data security regulations specific to different regions.

In addition, we find that 10% of privacy policies are not written in HTML format which affects the coverage of the section-level completeness analysis because our work relies on the HTML tags to classify the components. Future work will explore a method that can be applied to any privacy policy format.

Third, when analyzing the timeliness, QuPer relies on Wayback Machine to trace the updates of privacy policies. We note that the snapshots captured by Wayback Machine may not fully represent the update frequency of a web page. Due to this, our assessment may underrate some privacy policies. Aiming to estimate such underrating, we have consulted with the developers of Wayback Machine. They explain that there are many factors that determine the frequency of their snapshot taking, but these factors cannot be disclosed. Our future work will be committed to building a monitor to obtain precise update information.

9 RELATED WORK

QuPer aims to analyze the quality of VPA app's privacy policies. It has been inspired by existing work that automatically interprets privacy policies and assesses the compliance between policies and actually app behaviors. In this section, we briefly summarize these two lines of research.

Privacy policy analysis. The first line of research focuses on analyzing privacy policies in various domains. Andow et al. [11] develop PolicyLint, which automatically detects contradictions from the statements of the Android app's privacy policy. It uses sentence-level NLP to understand the declared statements on data collection and sharing. Benefiting from the substantial population of Android apps, it can derive precise ontologies of data objects and entities in the Android context. Liao et al. [48] conduct an empirical analysis of VPA apps' privacy policies, based on PolicyLint and their developed NLP-based methods. Their results reveal the quality issues in availability and disclosure, which highlight the necessity to develop a systematic approach to measuring privacy policy quality for VPA apps. Harkous et al. [37] develop Polisis to generate user-friendly summaries for various aspects of a given privacy policy, to aid users in efficiently comprehending the vast amount of privacy policy contents. It uses deep learning techniques on a large corpus for this purpose. Liu et al. [51] conduct a completeness checking against ten groups of essential information mandated by GDPR Art. 13. QuPer has been inspired by these studies in its several aspects, and it integrates all quality concerns of privacy policies collected from them to ensure the quality assessment is conducted comprehensively. This includes the measurement of readability and timeliness, which have not been well studied by previous work.

Consistency between privacy policies and data practices. Another line of research conducts consistency checking between privacy policies and actual behaviors. Andow et al. [12] propose POLICHECK based on PolicyLint [11] and AppCensus [4], to check the entity-sensitive consistency. Lentzsch et al. [46] conduct the first worldwide large-scale analysis of Alexa skills, focusing on the skill certification process. They also examine the consistency between skill privacy policies and their actual behaviors, and find that skills in the "kids" category exhibit the most severe violations, which necessitates the inclusion of the CHILDREN category in our compliance checking. Manandhar et al. [53] conduct an empirical large-scale analysis of smart home devices. They focus on examining the availability and coverage of privacy policies, aiming to gain insights into the current state of privacy disclosure within the smart home ecosystem. Xie et al. [81] develop Skipper to detect the noncompliance between skills' behaviors and their declared profile. QuPer identifies emerging concerns such as children- and region-specific policies, and includes them in the taxonomy to enhance the completeness of its assessment. It also proposes corresponding assessment techniques within each quality metrics, taking into consideration the challenges in the VPA context. For example, due to the lack of a large-scale corpus, it uses a two-level classification (see Section 4.1).

10 CONCLUSION

In this work, we conduct the first systematic study on the quality of privacy policies in the VPA app domain. We develop QuPer, which aims to automatically extract the meta features and lingual features, and assesses the privacy policy quality of VPA apps (i.e., Amazon Alexa skills) based on them. QuPer proposes four quality metrics to measure different aspects of the VPA privacy policy quality and uniquely develops a two-step document processing method to analyze VPA privacy policy documents. Our work reveals a concerning state of current VPA privacy policy quality and raises an alert to the VPA app developers. We therefore encourage store operators to set up regulatory mechanisms to ensure the high standards of VPA privacy policies.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their helpful comments. This research is partially supported by the University of Queensland under the Cyber Seed Funding and NSRSG grant 4018264-617225.

REFERENCES

[1] 2014. *WaybackMachine.* Retrieved September 8, 2023 from https://archive.org/web/
[2] 2020. *Beautiful Soup Documentations.* Retrieved September 8, 2023 from https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] 2020. *SpaCy Documentations.* Retrieved September 8, 2023 from https://spacy.io
[4] 2021. *AppCensus.* Retrieved September 8, 2023 from https://www.appcensus.io/search
[5] 2022. *Selenium website.* Retrieved September 8, 2023 from https://www.selenium.dev/
[6] Wasi Uddin Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for privacy policies. *arXiv preprint arXiv:2101.00123* (2021).
[7] Wejdan Ahmed Almunive. 2020. *The design and development of Guidelines for interactive course organizers.* Ph. D. Dissertation. Virginia Tech.
[8] Amazon. 2022. *Alexa.* Retrieved September 8, 2023 from https://www.amazon.com/alexa-skills/b?ie=UTF8&node=13727921011
[9] Amazon. 2022. *Configure Permissions for Customer Information in Your Skill.* Retrieved September 8, 2023 from https://developer.amazon.com/en-US/docs/alexa/custom-skills/configure-permissions-for-customer-information-in-your-skill.html
[10] Amazon. 2022. *Security Testing for an Alexa Skill.* Retrieved September 8, 2023 from https://developer.amazon.com/en-US/docs/alexa/custom-skills/security-testing-for-an-alexa-skill.html
[11] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play.. In *USENIX Security Symposium.* 585–602.
[12] Benjami Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck. In *29st USENIX Security Symposium (USENIX Security 20).* 985–1002.
[13] Peter Breese and William Burman. 2005. Readability of notice of privacy forms used by major health care institutions. *Jama* 293, 13 (2005), 1588–1594.
[14] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language* 109 (2019), 104047.
[15] Rochelle A Cadogan et al. 2004. An imbalance of power: the readability of internet privacy policies. *Journal of Business & Economics Research (JBER)* 2, 3 (2004).
[16] Baiqi Chen, Tingmin Wu, Yanjun Zhang, Mohan Baruwal Chhetri, and Guangdong Bai. 2023. Investigating Users' Understanding of Privacy Policies of Virtual Personal Assistant Applications. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security.* 65–79.
[17] Wei-Jie Chen, Yuan-Hai Shao, Chun-Na Li, and Nai-Yang Deng. 2016. MLTSVM: A novel twin support vector machine to multi-label learning. *Pattern Recognition* 52 (2016), 61–74.
[18] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. 2020. Dangerous skills got certified: Measuring the trustworthiness of skill certification in voice personal assistant platforms. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security.* 1699–1716.
[19] Jack Corbett and Erisa Karafili. 2021. *Private Data Harvesting on Alexa Using Third-Party Skills.* 127–142. https://doi.org/10.1007/978-3-030-93747-8_9
[20] Data Privacy Manager. 2021. *Luxembourg DPA issues €746 Million GDPR Fine to Amazon.* Retrieved September 8, 2023 from https://dataprivacymanager.net/luxembourg-dpa-issues-e746-million-gdpr-fine-to-amazon/
[21] Lydia de la Torre. 2018. A guide to the california consumer privacy act of 2018. *Available at SSRN 3275571* (2018).
[22] Aaron Drapkin. 2022. *Data Breaches That Have Happened in 2022.* Retrieved September 8, 2023 from https://tech.co/news/data-breaches-2022-so-far
[23] Derar Eleyan, Abed Othman, and Amna Eleyan. 2020. Enhancing software comments readability using flesch reading ease score. *Information* 11, 9 (2020), 430.
[24] Adam EM Eltorai, Syed S Naqvi, Soha Ghanian, Craig P Eberson, Arnold-Peter C Weiss, Christopher T Born, and Alan H Daniels. 2015. Readability of invasive procedure consent forms. *Clinical and translational science* 8, 6 (2015), 830–833.
[25] Enchanted Learning. 2022. *Negative Vocabulary Word List.* https://www.enchantedlearning.com/wordlist/negativewords.shtml
[26] ESL Forums. 2022. *List of 30+ Useful Contrast Transition Words.* Retrieved September 8, 2023 from https://eslforums.com/contrast-transition-words/
[27] European Commission. 2018. *Guidelines on Transparency under Regulation 2016/679.* Retrieved September 8, 2023 from https://ec.europa.eu/newsroom/article29/items/622227
[28] European Data Protection Board. 2018. *Article 29 Working Party Guidelines on transparency under Regulation 2016/679.* Retrieved September 8, 2023 from https://edpb.europa.eu/about-edpb/more-about-edpb/legacy-art-29-working-party_en
[29] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the international conference on web intelligence.* 18–25.
[30] Daniel B Felker et al. 1981. Guidelines for Document Designers. (1981).
[31] Chaim Gartenberg. 2018. *Amazon finally opens up Alexa to developers to make money off third-party skills.* Retrieved September 8, 2023 from https://www.theverge.com/2018/5/3/17314916/amazon-alexa-skills-developers-third-party-in-app-purchases
[32] Google. 2022. *Meet your Google Assistant.* Retrieved September 8, 2023 from https://assistant.google.com/
[33] Google. 2023. *Provide information for Google Play's Data safety section.* https://support.google.com/googleplay/android-developer/answer/10787469?hl=en
[34] Niharika Guntamukkala, Rozita Dara, and Gary Grewal. 2015. A machine-learning based approach for measuring the completeness of online privacy policies. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA).* IEEE, 289–294.
[35] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. Skillexplorer: Understanding the behavior of skills in large scale. In *29st USENIX Security Symposium (USENIX*

Security 20). 2649–2666.
[36] Lieve Hamers et al. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management* 25, 3 (1989), 315–18.
[37] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27st USENIX Security Symposium (USENIX Security 18)*. 531–548.
[38] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.
[39] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.
[40] Intersoft Consulting. 2020. *General Data Protection Regulation(GDPR)*. Retrieved September 8, 2023 from https://gdpr-info.eu/
[41] Intersoft Consulting. 2021. *Art. 13 GDPR Information to be provided where personal data are collected from the data subject*. Retrieved September 8, 2023 from https://gdpr-info.eu/art-13-gdpr/
[42] Muhammad Usama Islam and Beenish Moalla Chaudhry. 2023. Learnability Assessment of Speech-Based Intelligent Personal Assistants by Older Adults. In *International Conference on Human-Computer Interaction*. Springer, 321–347.
[43] Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 471–478.
[44] Bret Kinsella. 2022. *Alexa Skill Counts sURPASS 80k in US, Spain Adds the Most Skills, New Skill Rate Falls Globally*. Retrieved September 8, 2023 from https://voicebot.ai/2021/01/14/alexa-skill-counts-surpass-80k-in-us-spain-adds-the-most-skills-new-skill-introduction-rate-continues-to-fall-across-countries/
[45] Deepti Lamba and William H Hsu. 2021. Answer-Agnostic Question Generation in Privacy Policy Domain using Sequence-to-Sequence and Transformer Models. In *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*. IEEE, 256–261.
[46] Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das, and William Enck. 2021. Hey Alexa, is this skill safe?: Taking a closer look at the Alexa skill ecosystem. *Network and Distributed Systems Security (NDSS) Symposium*.
[47] Suwan Li, Lei Bu, Guangdong Bai, Zhixiu Guo, Kai Chen, and Hanlin Wei. 2022. VITAS: Guided Model-Based VUI Testing of VPA Apps. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*.
[48] Song Liao, Christin Wilson, Cheng Long, Hongxin Hu, and Huixing Deng. 2021. Problematic Privacy Policies of Voice Assistant Applications. *IEEE Security & Privacy* 19, 6 (2021), 66–73.
[49] Lorelei Lingard and Christopher Watling. 2021. Mastering the sentence. In *Story, Not Study: 30 Brief Lessons to Inspire Health Researchers as Writers*. Springer, 77–81.
[50] Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *2016 AAAI Fall Symposium Series*.
[51] Shuang Liu, Baiyang Zhao, Renjie Guo, Guozhu Meng, Fan Zhang, and Meishan Zhang. 2021. Have you been properly notified? automatic compliance analysis of privacy policy text with GDPR article 13. In *Proceedings of the Web Conference*. 2154–2164.
[52] Kulani Mahadewa, Yanjun Zhang, Guangdong Bai, Lei Bu, Zhiqiang Zuo, Dileepa Fernando, Zhenkai Liang, and Jin Song Dong. 2021. Identifying privacy weaknesses from multi-party trigger-action integration platforms. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 2–15.
[53] Sunil Manandhar, Kaushal Kafle, Benjamin Andow, Kapil Singh, and Adwait Nadkarni. 2022. Smart Home Privacy Policies Demystified: A Study of Availability, Content, and Coverage. In *31st USENIX Security Symposium (USENIX Security 22)*. 3521–3538.
[54] Aleecia McDonald, Robert Reeder, Patrick Kelley, and Lorrie Cranor. 2009. A Comparative Study of Online Privacy Policies and Formats. 37–55. https://doi.org/10.1007/978-3-642-03168-7_3
[55] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.
[56] Aleecia M McDonald, Robert W Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 37–55.
[57] Anca Micheti, Jacquelyn Burkell, and Valerie Steeves. 2010. Fixing broken doors: Strategies for drafting privacy policies young people can understand. *Bulletin of Science, Technology & Society* 30, 2 (2010), 130–143.
[58] George Milne and Mary Culnan. 2004. Strategies for Reducing Online Privacy Risks: Why Consumers Read (Or Don't Read) Online Privacy Notices. *Journal of Interactive Marketing* 18 (06 2004), 15 – 29. https://doi.org/10.1002/dir.20009
[59] Office of the Attorney General. 2022. *California Consumer Privacy Act (CCPA)*. Retrieved September 8, 2023 from https://oag.ca.gov/privacy/ccpa
[60] Gayatri Prajapati and Khwairakpam Sharmila. 2023. Difficulties experienced by older adults when not using assistive devices. *Discover Social Science and Health* 3, 1 (2023), 8.
[61] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4947–4958.
[62] Readable. 2023. *The Automated Readability Index*. Retrieved September 8, 2023 from https://readable.com/readability/automated-readability-index
[63] Readable. 2023. *Flesch Reading Ease and the Flesch Kincaid Grade Level*. Retrieved September 8, 2023 from https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/
[64] Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, et al. 2019. Comparing and developing tools to measure the readability of domain-specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4831–4842.
[65] Jason DM Rennie. 2001. Improving multi-class text classification with naive Bayes. (2001).
[66] Elliot Schumacher and Maxine Eskenazi. 2016. A readability analysis of campaign speeches from the 2016 US presidential campaign. *arXiv preprint arXiv:1603.05739* (2016).
[67] RJ Senter and Edgar A Smith. 1967. *Automated readability index*. Technical Report. Cincinnati Univ OH.
[68] Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*. 574–576.
[69] Bengt Sigurd, Mats Eeg-Olofsson, and Joost Van Weijer. 2004. Word length, sentence length and frequency–Zipf revisited. *Studia linguistica* 58, 1 (2004), 37–52.
[70] van der Hof Simone and Eva Lievens. 2018. The importance of privacy by design and data protection impact assessments in strengthening protection of children's personal data under the GDPR. *Communications law* 23, 1 (2018).
[71] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. 1 (2021).
[72] Peter Story, Sebastian Zimmeck, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Natural language processing for mobile app privacy compliance. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.
[73] Lior Jacob Strahilevitz and Matthew B Kugler. 2016. Is privacy policy language irrelevant to consumers? *The Journal of Legal Studies* 45, S2 (2016), S69–S95.
[74] Ruoxi Sun and Minhui Xue. 2020. Quality assessment of online automated privacy policy generators: an empirical study. In *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*. 270–275.
[75] Elizabeth Tebeaux and Linda Driskill. 2020. Culture and the shape of rhetoric: Protocols of international document design. In *Exploring the rhetoric of international professional communication*. Routledge, 211–251.
[76] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. 15–21.
[77] Bergur Thormundsson. 2023. *Virtual Assistant Technology - statistics & facts*. Retrieved September 8, 2023 from https://www.statista.com/topics/5572/virtual-assistants
[78] Pauline Velez and Steven D Ashworth. 2007. The impact of item readability on the endorsement of the midpoint response in surveys. In *Survey Research Methods*, Vol. 1. 69–74.
[79] Payton Walker and Nitesh Saxena. 2021. Evaluating the Effectiveness of Protection Jamming Devices in Mitigating Smart Speaker Eavesdropping Attacks Using Gaussian White Noise. In *Annual Computer Security Applications Conference*. 414–424.
[80] Fuman Xie, Yanjun Zhang, Hanlin Wei, and Guangdong Bai. 2022. UQ-AAS21: A Comprehensive Dataset of Amazon Alexa Skills. In *Advanced Data Mining and Applications (ADMA)*. 159–173.
[81] Fuman Xie, Yanjun Zhang, Chuan Yan, Suwan Li, Lei Bu, Kai Chen, Zi Huang, and Guangdong Bai. 2022. Scrutinizing Privacy Policy Compliance of Virtual Personal Assistant Apps. In *37th IEEE/ACM International Conference on Automated Software Engineering*.
[82] Jeffrey Young, Song Liao, Long Cheng, Hongxin Hu, and Huixing Deng. 2022. SkillDetective: Automated Policy-Violation Detection of Voice Assistant Applications in the Wild. In *31st USENIX Security Symposium (USENIX Security 22)*.
[83] Ke Zhou, Miriam Redi, Andrew Haines, and Mounia Lalmas. 2016. Predicting pre-click quality for native advertisements. In *Proceedings of the 25th International Conference on World Wide Web*. 299–310.

A STATUS CODE DETAILS

Table 10 provides the details of the status codes returned from the website mentioned in Section 6.1.

Table 10: Status code details

Code	Status code name	Description	Code	Status code name	Description
200	OK	The request is successful.	406	Not Acceptable	The server cannot complete the request.
400	Bad Request	Client requests syntax error.	408	Timeout	The request timeout.
401	Unauthorized	The request requires the user's identity.	500	Internal Server Error	The server has an internal error.
403	Forbidden	The server refuses to execute the request.	503	Service Unavailable	The server is Unavailable.
404	Not Found	The server cannot find the resource.	999	Certificate Error	The SSH certificate is incorrectly configured.

B OVERVIEW OF READABILITY METRICS

Table 11 explains the readability metrics discussed in Section 7.1.

Table 11: Overview of Readability Metrics

Metric	Description	Scope Mapping
Automated Readability Index (ARI)	The Automated Readability Index (ARI) is used by U.S. military to assess the grade level to read text [62], and has also been widely used in various studies [24, 29] to measure the readability of articles and paragraphs. It calculates a numerical score based on factors such as sentence length and word difficulty, providing valuable insights into the reading level required to comprehend a particular piece of text.	1 = Kindergarten, 2 = First Grade, 3 = Second Grade, 4 = Third Grade, 5 = Fourth Grade, 6 = Fifth Grade, 7 = Sixth Grade, 8 = Seventh Grade, 9 = Eighth Grade, 10 = Ninth Grade, 11 = Tenth Grade, 12 = Eleventh Grade, 13 = Twelfth Grade, 14 = College student
Flesch Readability Ease Score (FRES)	Flesch Reading Ease score (FRES) [63] is widely adopted by individuals and organizations seeking to ensure that their written content is accessible and easily understandable to their target audience [13, 23]. It provides an indication of how easy or difficult a text is to understand by considering both sentence-level complexity and word-level complexity.	90-100 = Very Easy, 80-89 = Easy, 70-79 = Fairly Easy, 60-69 = Standard, 50-59 = Fairly Difficult, 30-49 = Difficult, 0-29 = Very Confusing
Laesbarheds Index (LIX)	Laesbarheds Index (LIX) is another metric that has been utilized to determine the difficulty of documentation [29, 83]. It measures a text's readability based on factors such as sentence length, complexity, unusual words, main words, and different words used. One of its advantages is that its reliability considers a wide range of age groups, from children's literature to adult reading materials.	0-29 = It is very easy to read, 30-39 = It is easy to read, 40-49 = It is a little hard to read, 50-59 = It is hard to read, 60 = It is very hard to read
Syllables per Word (SPW)	"Syllables per word" is a measure that calculates the average number of syllables in each word. It indicates the complexity of words in a given context.	1.5 = Second Grade, 1.6 = Third to Eighth Grade, Adult reading average SPW = 1.77
Word per Sentence (WPS)	"Words per sentence" refers to the average number of words in each sentence. It helps determine the length and complexity of sentences in a given text.	10.6 = Second Grade, 13.9 = Third to Fifth Grade, 14.7 = Sixth to Eighth Grade, Adult reading average WPS = 15.24
Letters per Word (LPW)	"Letters per word" measures the average number of letters in each word and provides insights into the word complexity and length in a given context.	Adult reading average LPW = 5.24
Sentence Count (SC)	"Sentence count" refers to the number of sentences present in a given text or context.	It is 238 words per minute for English silent reading and 183 words per minute for speaking [14]. Privacy policies should be tailored to align with users' reading habits to determine the length and content of the document.
Word Count (WC)	"Word count" refers to the total number of words present in a given text or context.	
Reading time (RT)	"Reading time" refers to the estimated time it takes for an individual to read a particular piece of text or content. It is a measurement used to provide readers with an estimate of how long it will take them to go through the material.	
Speaking time (ST)	The time it would take for the average person to say this text aloud at a rate of 125 words per minute.	

C AN INSTRUCTION TO GUIDE THE ORDERING OF LOGICAL SEQUENCE OF SECTIONS

Table 12 provides a tutorial on the sequencing of privacy policy sections discussed in Section 7.1. We also ask the participants to briefly explain their reasons of the ordering, and some examples are summarized in Table 13.

Table 12: Logical sequence of section guidelines tutorial

Section Sorting Tutorial		
1. Design Purpose: Assisting volunteers in understanding the four guidelines of section sorting and helping volunteers to complete the section sorting task. **2. Audience: Volunteers with experience writing legal documents or in the computer field.** **3. Reference: Felker, Daniel B. "Guidelines for Document Designers." (1981).**		
Guideline	**Explanation**	**Example**
Put content in a time sequence	One effective approach to presenting time-related information is to arrange sentences and paragraphs in a chronological sequence. Time-related information consists of actions or events that are interdependent and rely on their occurrence in a specific order. In procedural instructions, steps are often interconnected based on their temporal relationship.	Consider the process of making a cup of tea. 1. Place the kettle on the stove and turn on the heat. 2. Once the water reaches a rolling boil, pour it into the cup. 3. Place the tea bag in the cup and let it steep for a few minutes. 4. Stir the tea to ensure it's well-mixed. 5. Allow the tea to cool for a moment, then enjoy your cup of tea.
Present the general information before the specific one	Prioritizing the "bigger picture" before delving into the individual elements is a recommended approach. This principle aligns with the longstanding relationship between a paragraph's topic sentence and the supporting sentences that provide further details and elaboration.	Instead of immediately diving into the specific features of a new smartphone, it is beneficial to first provide an overview of its overall capabilities and functionalities. By presenting the "big picture" of the smartphone's capabilities, such as its advanced camera system, high-resolution display, and powerful processing capabilities, readers can better understand the significance and value of the subsequent detailed discussions on individual features and specifications.
Discuss things that affect many people before those that affect few	When writing public documents, it is common for the topics to have relevance to a wide audience. If the information you are presenting applies to the majority of the group, while some information is specific to only a few individuals, it is advisable to prioritize the discussion of the information that is relevant to the largest number of people.	Let's say you are writing a public document about a new transportation initiative in a city. The document covers various aspects, including public transit improvements and bike lane expansions. Given that public transit impacts a larger portion of the population compared to bike lanes, it would be logical to discuss the public transit improvements first in the document.
Present permanent provisions before temporary ones	If you are writing content that includes provisions with long-term implications as well as others that are applicable only under specific conditions or at certain times, it is advisable to present the long-term content first.	Assuming you are writing a cookbook, it is advisable to present the long-term applicable cooking techniques and common ingredient uses first. This way, you establish a foundation of basic cooking knowledge and skills for the readers. These fundamental techniques may include methods for vegetable preparation and essential principles for cooking meat. By introducing these techniques initially, you equip the readers with the necessary skills to achieve good cooking results in their everyday culinary endeavors. Subsequently, you can provide recipes for special occasions or festive dishes, such as Christmas dinner recipes or birthday cake preparations.

Table 13: Feedback of participants on their reasons of the ordering

Guideline	**Feedback**
Put content in a time sequence	1. "*Based on this guideline, I believe that the* COLLECT *section should be placed first because data collection is a prerequisite for data protection, data usage, and other related aspects.* " 2. "COOKIE *section should be placed towards the beginning because, typically, when we visit websites or use apps, we are often prompted with the question of whether we accept cookies first.* "
Present the general information before the specific one	1. "*I think the process of data handling falls under the category of "general information" as mentioned in this guideline. This includes aspects such as data collection, data sharing, data usage, and data security.*" 2. "*In my opinion,* PROVIDER *section should be categorized as "specific information" and placed toward the latter part of the privacy policy.*"
Discuss things that affect many people before those that affect few	1. "CHILDREN *section should be placed towards the later part as it pertains to a smaller subset of the population.*" 2. "*I believe that the* REGION *section should be placed towards the later part because the regulations mentioned within it apply only to a specific region or group of people.*" 3. "*I think there should be a reference to the rights of people in special areas in* REGION*, and that should be put after* RIGHT *section.*" 4. "*I think* CHILDREN *section should be put in front of* REGION *section because, for some products targeting kids, children are the main audience.*"